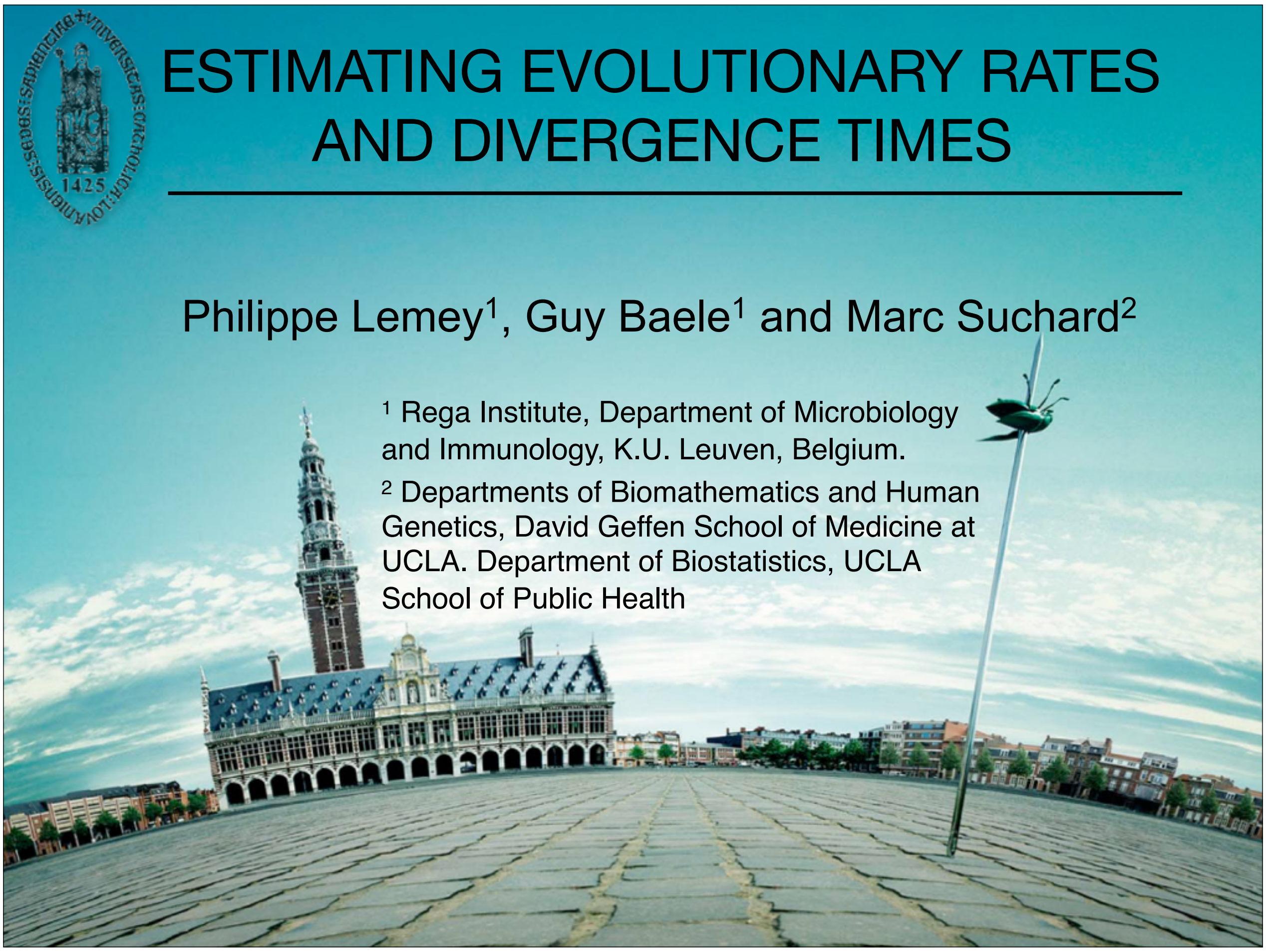
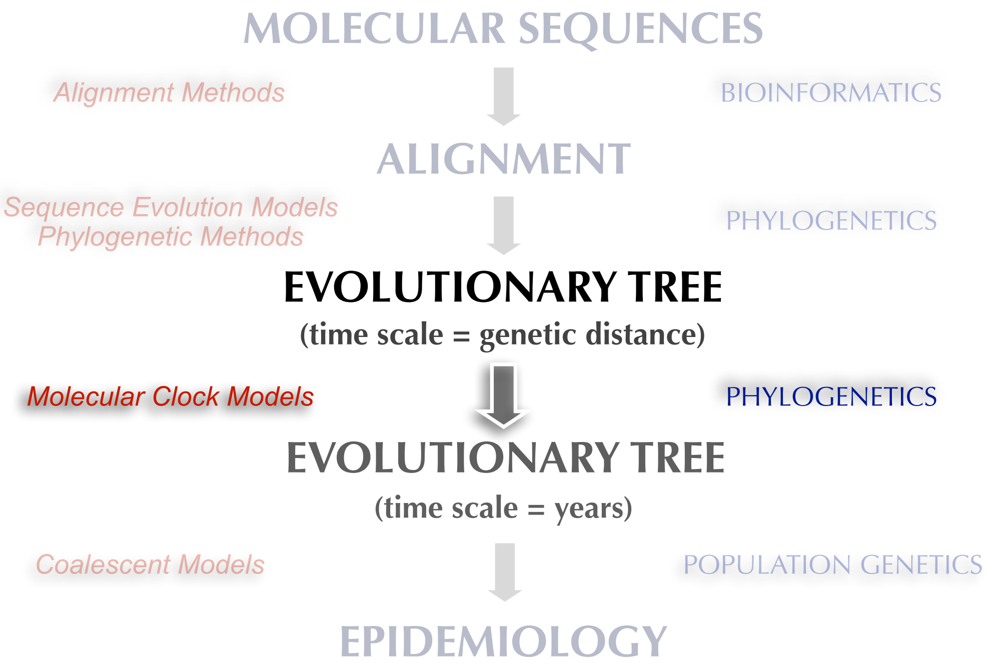


ESTIMATING EVOLUTIONARY RATES AND DIVERGENCE TIMES

Philippe Lemey[1], Guy Baele[1] and Marc Suchard[2]

[1] Rega Institute, Department of Microbiology and Immunology, K.U. Leuven, Belgium.

[2] Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA. Department of Biostatistics, UCLA School of Public Health



MOLECULAR SEQUENCES

Alignment Methods

BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods

PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

Molecular Clock Models

PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = years)

Coalescent Models

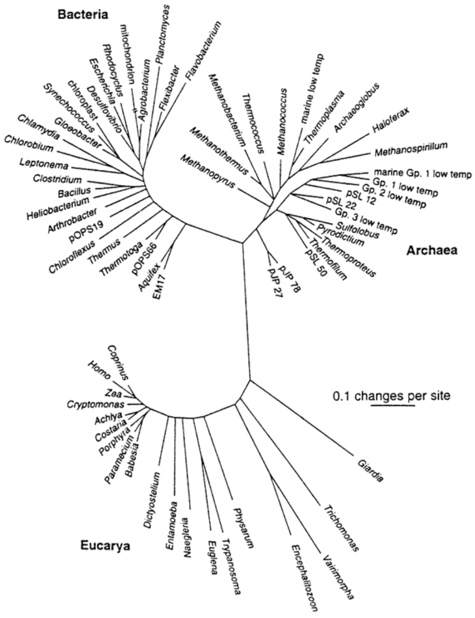
POPULATION GENETICS

EPIDEMIOLOGY

Molecular phylogenies

- most molecular phylogenies
  - are unrooted (or the rooting is due to prior information)
  - have branch lengths representing genetic change

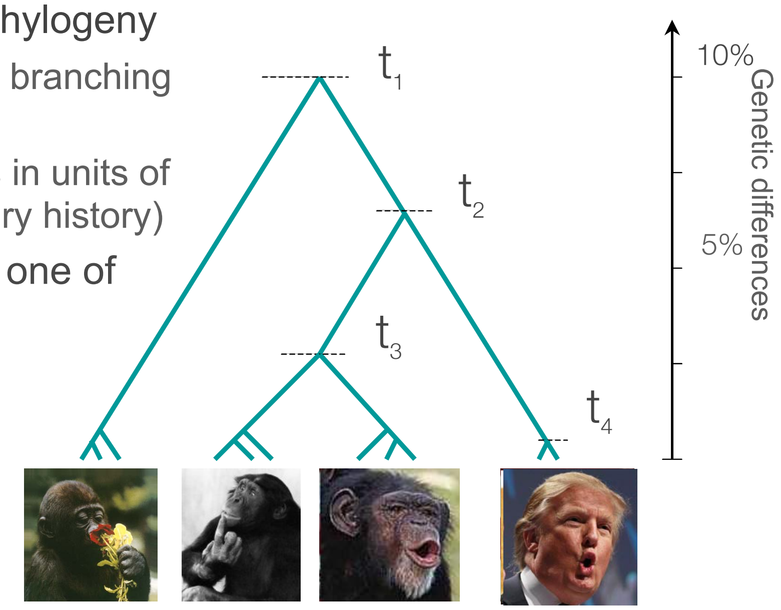
The Tree of Life



- NR Pace, Science 1997

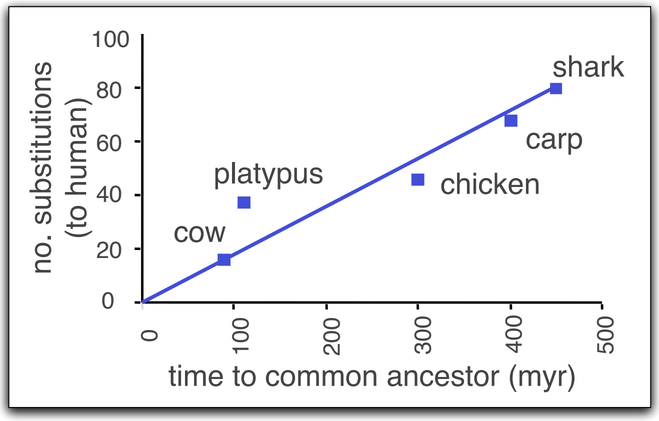
Molecular phylogenies

- the ideal molecular phylogeny
  - is rooted (implies a branching order)
  - has branch lengths in units of time (an evolutionary history)
- how do we construct one of these trees?



A constant evolutionary rate through time

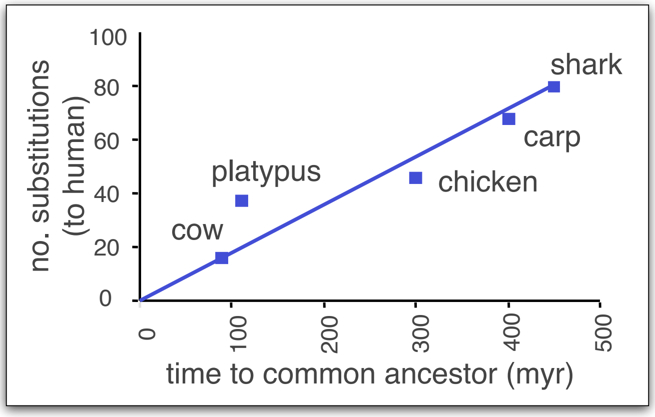
- to obtain a time phylogeny, the evolutionary model must assume a relationship between the accumulation of genetic diversity and time



- Zuckerkandl and Pauling (1962): the rate of amino acid replacements in animal haemoglobins was roughly proportional to real time, as judged against the fossil record

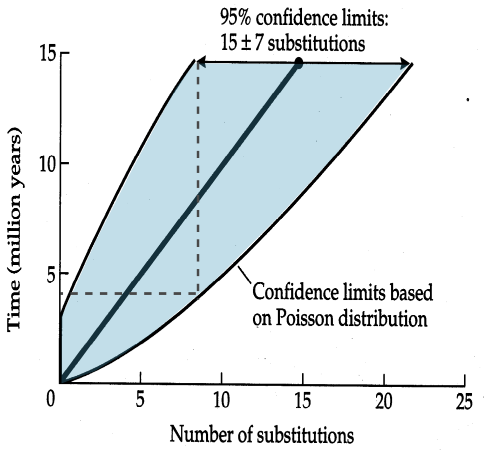
A constant evolutionary rate through time

- the *molecular clock* is particularly striking when compared to the obvious differences in rates of morphological evolution...



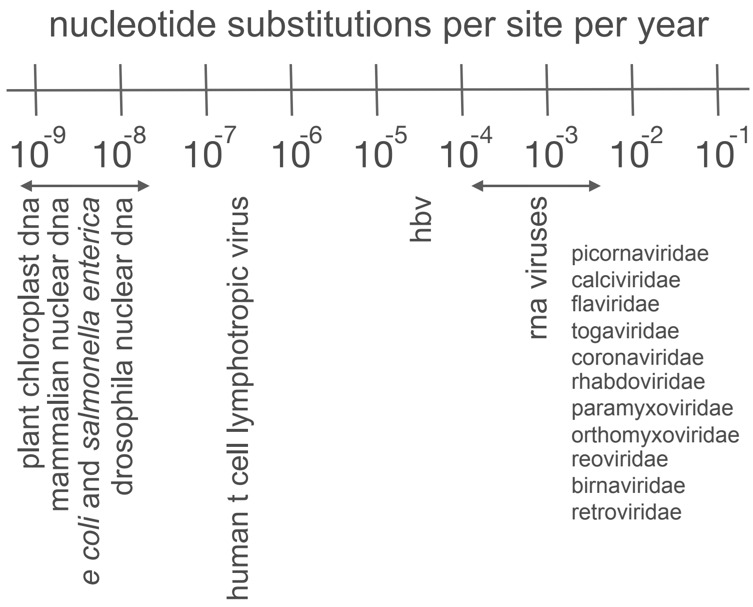
The molecular clock is not a metronome

- if mutation every MY with Poisson variance
  - 95% of the lineages 15MY old have 8-22 substitutions
  - 8 substitutions also could be < 5 MY old



- Molecular Systematics, p532.

And there is no global molecular clock



nucleotide substitutions per site per year

10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}

plant chloroplast dna
mammalian nuclear dna
e coli and *salmonella enterica*
drosophila nuclear dna

human t cell lymphotropic virus

hbv

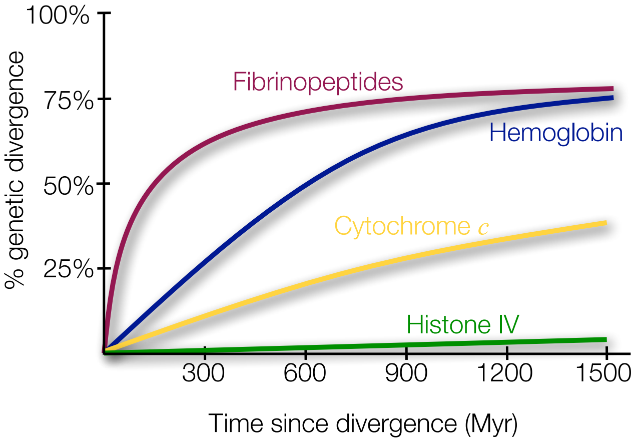
rna viruses

picornaviridae
calciviridae
flaviridae
togaviridae
coronaviridae
rhabdoviridae
paramyxoviridae
orthomyxoviridae
reoviridae
birnaviridae
retroviridae

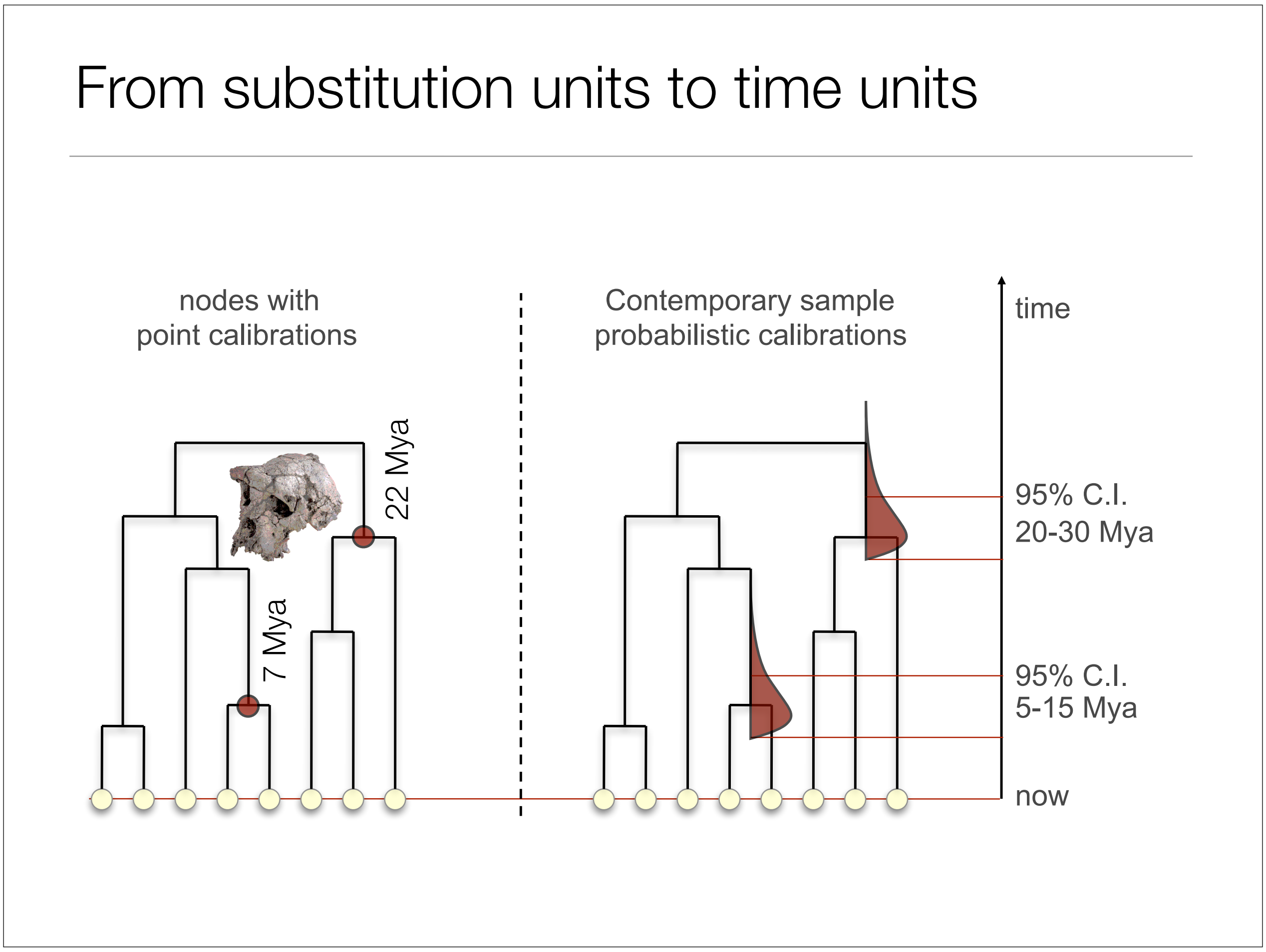
And there is no global molecular clock

- different genes, different profiles
- variation in mutation rate?
- variation in selection

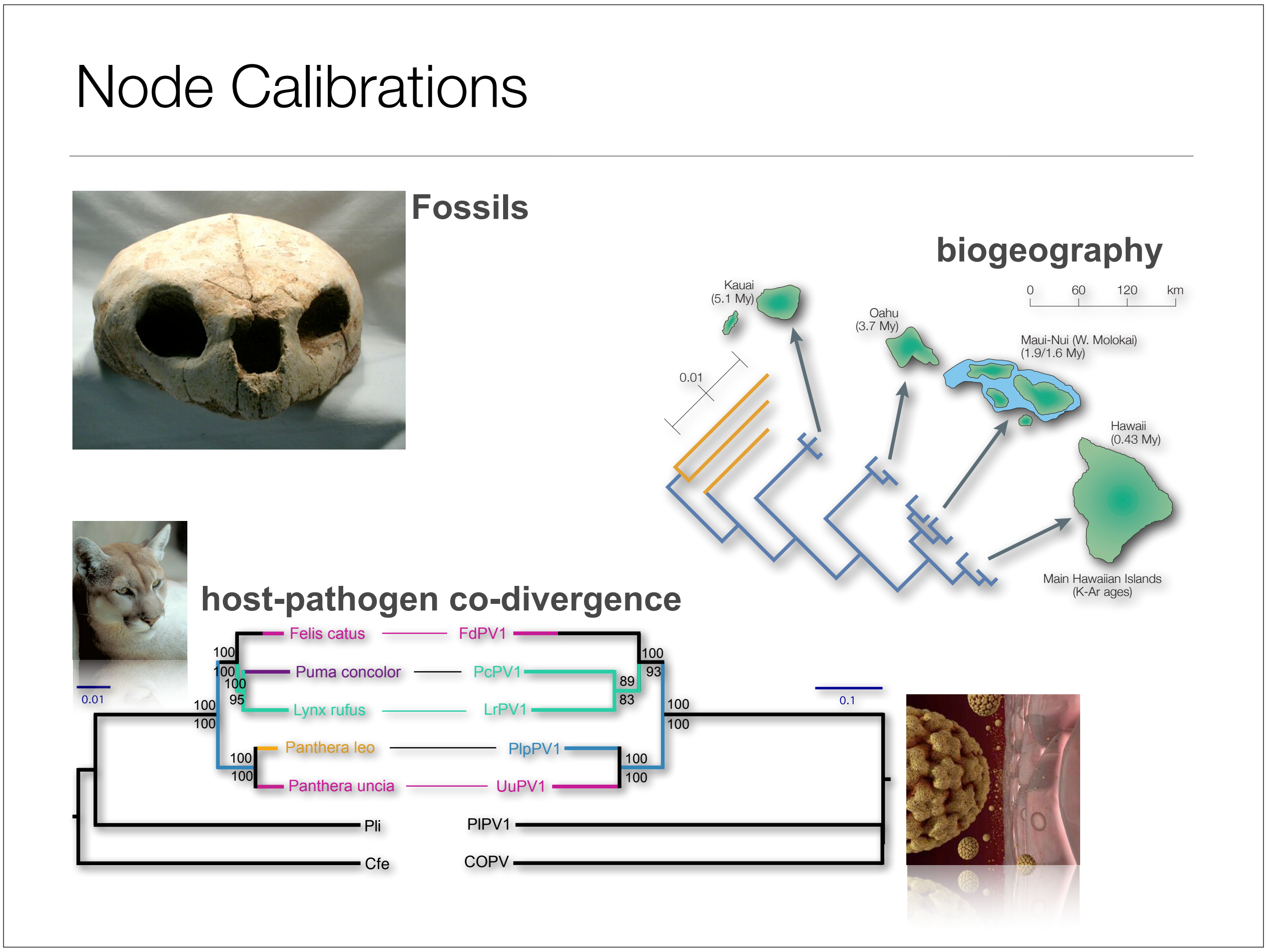
  genes coding for some molecules under very strong stabilizing selection



calibrating the molecular clock

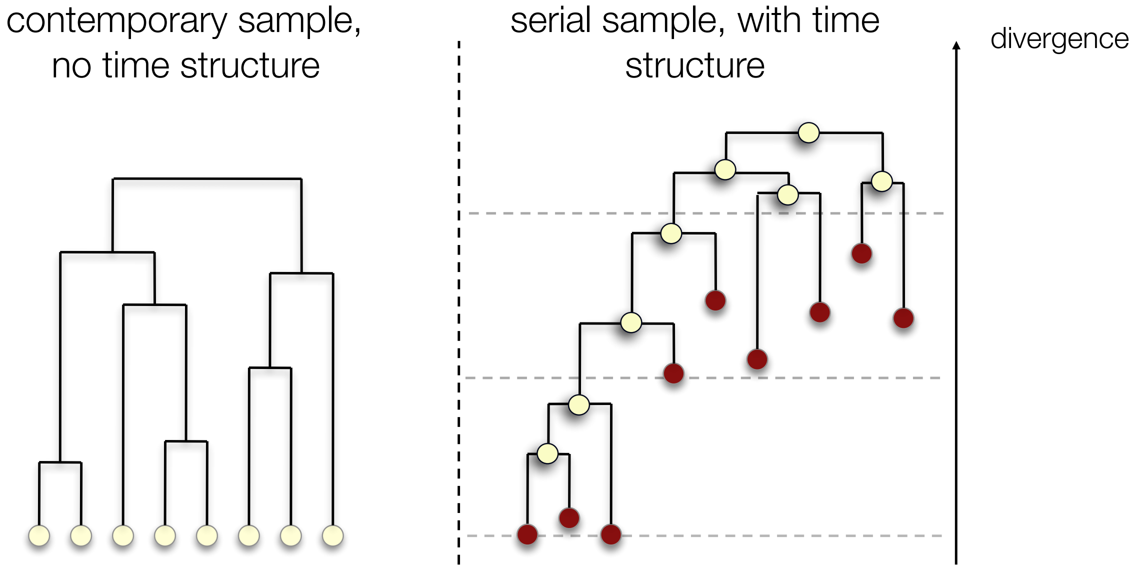


From substitution units to time units



Node Calibrations

Calibration using sampling times

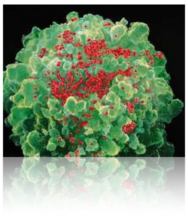


contemporary sample, no time structure

serial sample, with time structure

divergence

Tip calibration: two major applications



RNA viruses

evolve quickly:

10^{-3} - 10^{-5} substitutions per site per year.

- Substitutions accumulate between the times of sampling
- Serially sampled sequences or heterochronous sequences

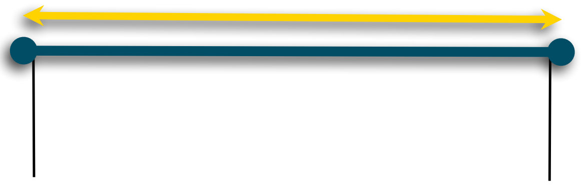
Measurably evolving population

ancient DNA

data sets of radiocarbon-dated specimens

incorporating sampling time: naive method

observed number of substitutions
or genetic divergence
d

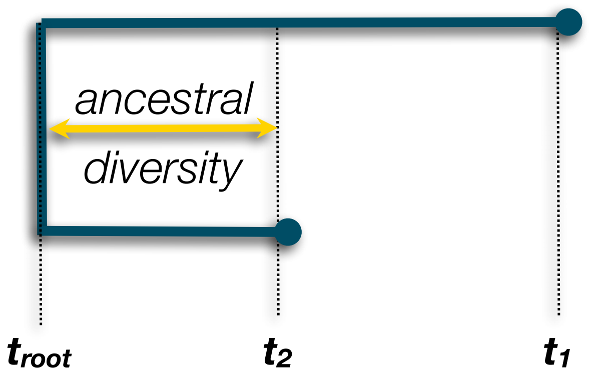


sampling time 1
t_1

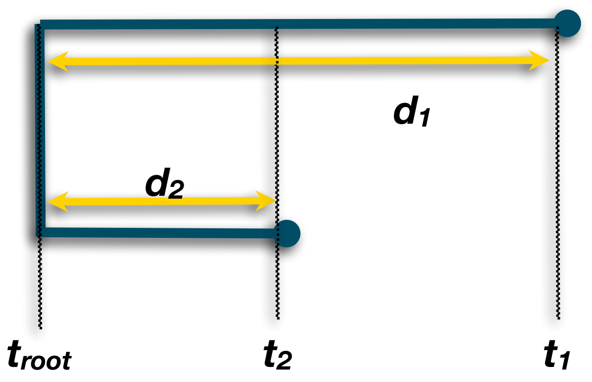
sampling time 2
t_2

substitution rate, μ
$= d / |t_1 - t_2|$

incorporating sampling time: naive method

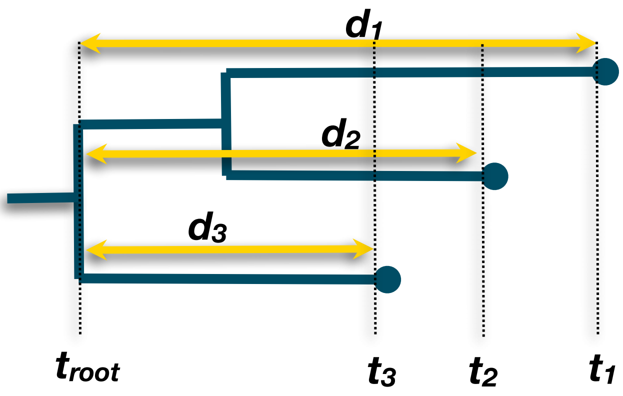


incorporating sampling time: naive method



$$\mu = (d_1 - d_2) / (t_1 - t_2)$$

linear regression



$$\mu = d_i / (t_i - t_{root})$$

- can be rearranged:

$$d_i = \mu \, (t_i - t_{root})$$

$$E[d_i] = \mu \, . \, t_i - \mu \, . \, t_{root}$$

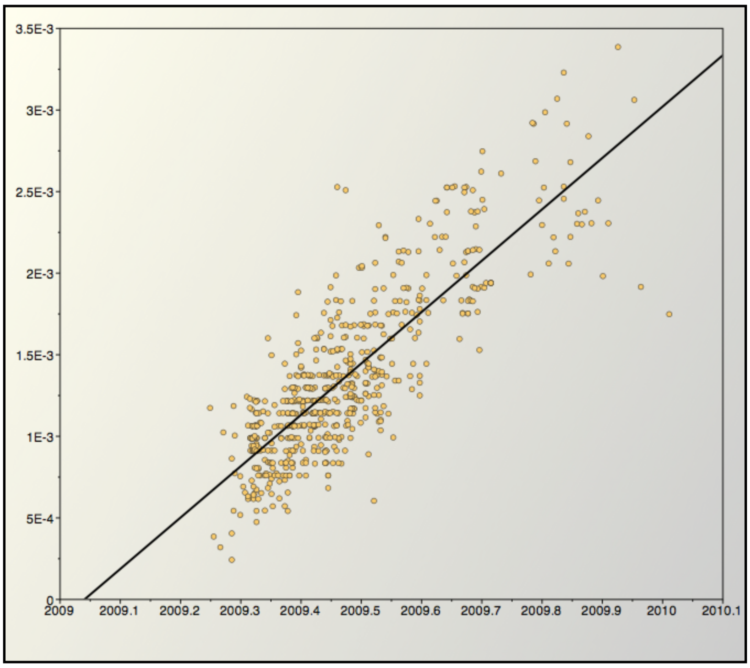
gradient is: μ

y-intercept is: $- \mu \, . \, t_{root}$

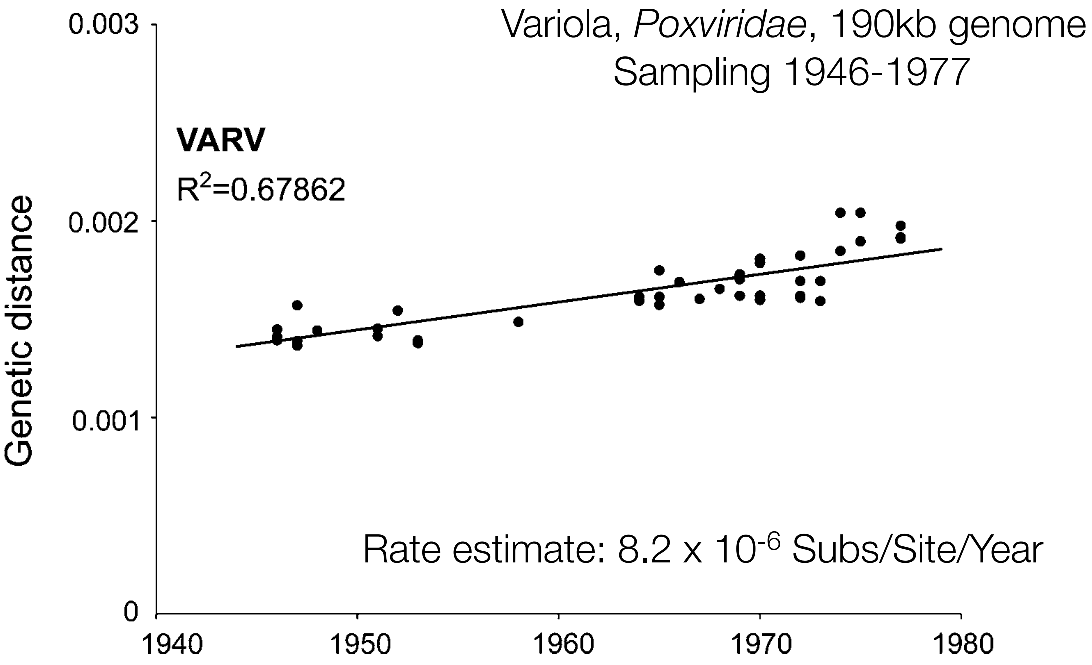
x-intercept is: t_{root}

Estimating the time-scale

- H1N1/09 'Swine Flu'
- Rate: 3.14E^{-3} mutations/genomic site/year
- tMRCA: 2009.041 (15-Jan-2009)
- Correlation: 0.83
- R^2: 0.69



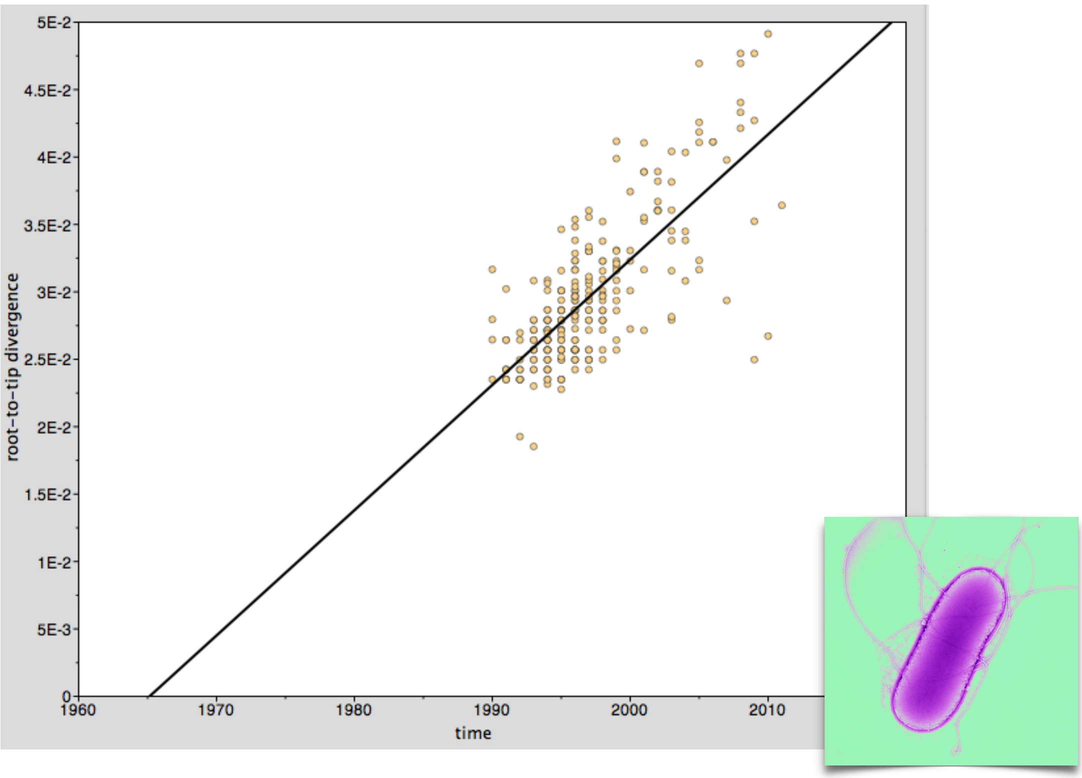
A DNA virus (smallpox)


Variola, *Poxviridae*, 190kb genome
Sampling 1946-1977

VARV
R^2=0.67862

Rate estimate: 8.2 x 10^{-6} Subs/Site/Year

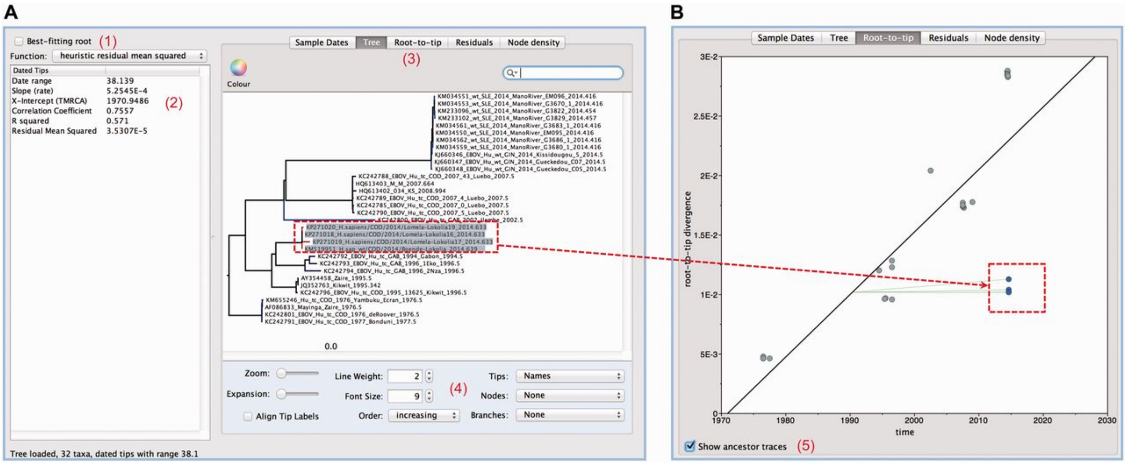
Salmonella Typhimurium



Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)

Andrew Rambaut, Tommy T. Lam, Luiz Max Carvalho, Oliver G. Pybus

DOI: http://dx.doi.org/10.1093/ve/vew007 vew007 First published online: 10 April 2016


A
(1)
(2)
Sample Dates
Tree
Root-to-tip
Residuals
Node density
(3)
Colour
KM034551_wt_SLE_2014_ManoRiver_EM096_2014.416
KM034553_wt_SLE_2014_ManoRiver_G3670_1_2014.416
KM233096_wt_SLE_2014_ManoRiver_G3822_2014.454
KM233102_wt_SLE_2014_ManoRiver_G3829_2014.457
KM034561_wt_SLE_2014_ManoRiver_G3683_1_2014.416
KM034550_wt_SLE_2014_ManoRiver_EM095_2014.416
KM034562_wt_SLE_2014_ManoRiver_G3686_1_2014.416
KM034559_wt_SLE_2014_ManoRiver_G3680_1_2014.416
KJ660346_EBOV_Hu_wt_GIN_2014_Kissidougou_5_2014.5
KJ660347_EBOV_Hu_wt_GIN_2014_Gueckedou_C07_2014.5
KJ660348_EBOV_Hu_wt_GIN_2014_Gueckedou_C05_2014.5
KC242788_EBOV_Hu_tc_COD_2007_43_Luebo_2007.5
HQ613403_M_M_2007.664
HQ613402_034_KS_2008.994
KC242789_EBOV_Hu_tc_COD_2007_4_Luebo_2007.5
KC242785_EBOV_Hu_tc_COD_2007_0_Luebo_2007.5
KC242790_EBOV_Hu_tc_COD_2007_5_Luebo_2007.5
KC242800_EBOV_Hu_tc_GAB_2002_Ilembe_2002.5
KP271020_H.sapiens/COD/2014/Lomela-Lokolia19_2014.633
KP271018_H.sapiens/COD/2014/Lomela-Lokolia16_2014.633
KP271019_H.sapiens/COD/2014/Lomela-Lokolia17_2014.633
KM519951_H.sap_wt/COD/2014/Boende-Lokolia_2014.639
KC242792_EBOV_Hu_tc_GAB_1994_Gabon_1994.5
KC242793_EBOV_Hu_tc_GAB_1996_1Eko_1996.5
KC242794_EBOV_Hu_tc_GAB_1996_2Nza_1996.5
AY354458_Zaire_1995.5
JQ352763_Kikwit_1995.342
KC242796_EBOV_Hu_tc_COD_1995_13625_Kikwit_1996.5
KM655246_Hu_tc_COD_1976_Yambuku_Ecran_1976.5
AF086833_Mayinga_Zaire_1976.5
KC242801_EBOV_Hu_tc_COD_1976_deRoover_1976.5
KC242791_EBOV_Hu_tc_COD_1977_Bonduni_1977.5
0.0
Zoom:
Expansion:
(4)
B
Sample Dates
Tree
Root-to-tip
Residuals
Node density
3E-2
2.5E-2
2E-2
1.5E-2
1E-2
5E-3
0
root-to-tip divergence
1970
1980
1990
2000
2010
2020
2030
time
(5)

□ Best-fitting root

Function: heuristic residual mean squared

Dated Tips	
Date range	38.139
Slope (rate)	5.2545E-4
X-Intercept (TMRCA)	1970.9486
Correlation Coefficient	0.7557
R squared	0.571
Residual Mean Squared	3.5307E-5

Line Weight: 2

Font Size: 9

□ Align Tip Labels

Order: increasing

Tips: Names

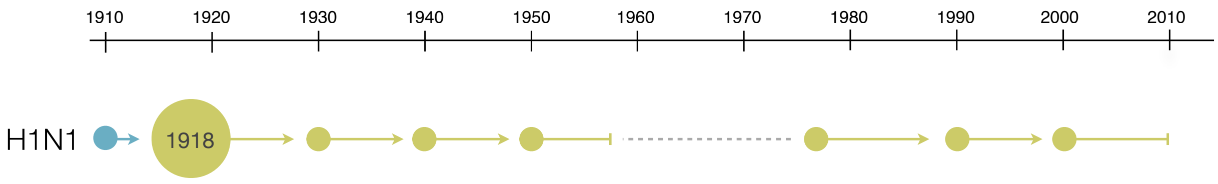
Nodes: None

Branches: None

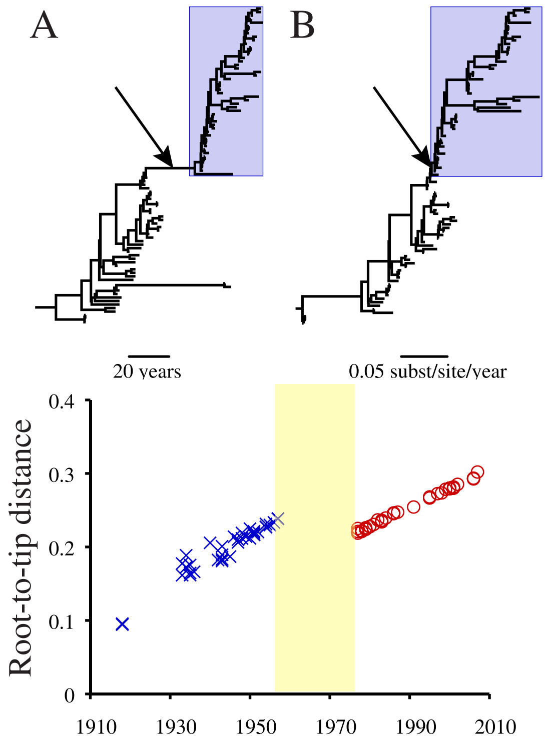
Tree loaded, 32 taxa, dated tips with range 38.1

☑ Show ancestor traces

Two lost decades of seasonal H1N1 evolution



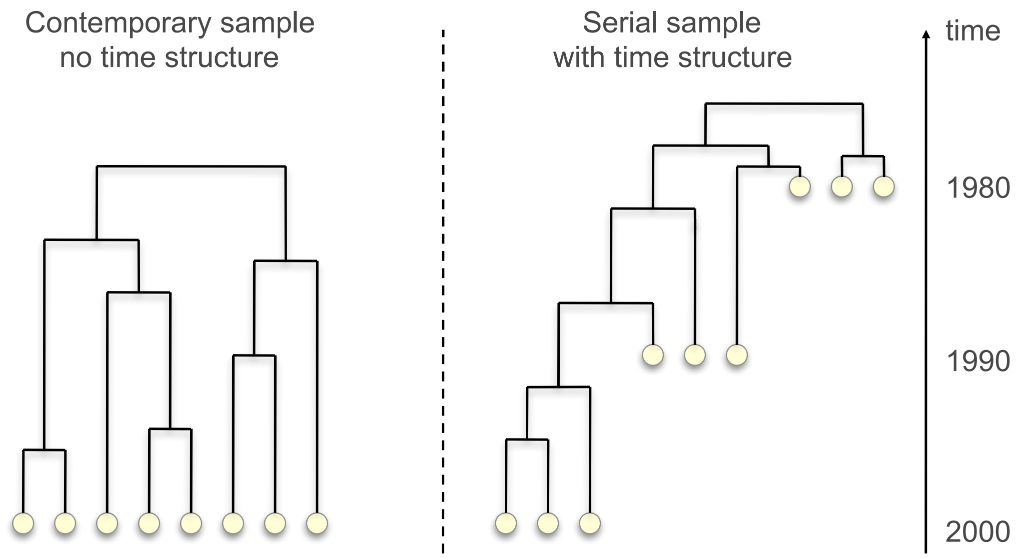
Two lost decades of seasonal H1N1 evolution



- Linear regression of genetic distance against sampling time shows no divergence between 1957 and 1977 (when H1N1 re-emerged).
- Also apparent when comparing molecular clock trees (A) with non-clock trees (B).
- H1N1 re-emergence is thought to have been an accidental lab release.

Wertheim (2010) PLoS One. 5:e11184

Time structure via tip calibration



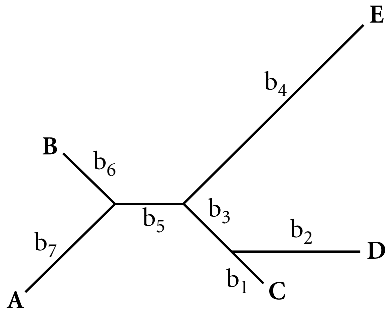
- Rambaut A. (2000) *Bioinformatics*, **16**, 395-399.

Clock versus non-clock

- strict molecular clock:
  Zuckerkandl & Pauling (1962) in Horizons in Biochemistry, pp. 189–225
  - all lineages evolve at the same rate
  - allows the estimation of the root of the tree and dates of individual nodes
- unconstrained (unrooted) Felsenstein model:
  Felsenstein (1981) *JME*, **17**: 368 - 376
  - each branch has its own rate independent of all others
  - time and rate are confounded and can only be estimated as a compound parameter (branch lengths)

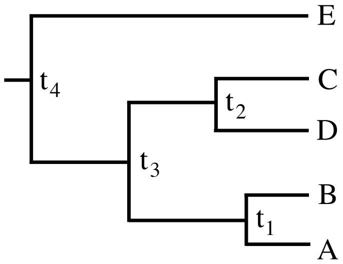
Likelihood ratio test for molecular clock models

- complex model H_1



2N-3 parameters

- null model H_0

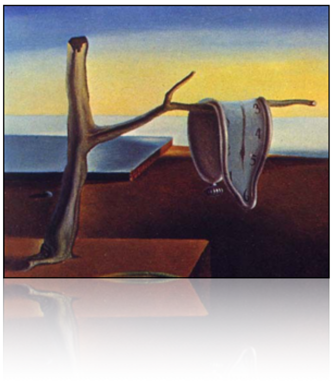


N-1 parameters

$LRS = 2(\max[\ln L(H_a|D)] - \max[\ln L(H_0|D)])$

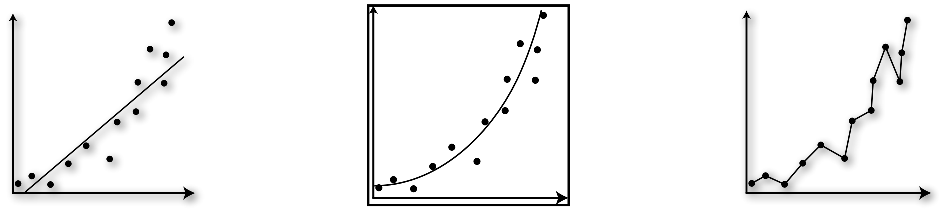
- likelihood ratio test with N-2 degrees of freedom
- models are nested because values of b_1-b_7 can be specified that give node heights t_1-t_4

Relaxing the molecular clock



Need for a relaxed molecular clock

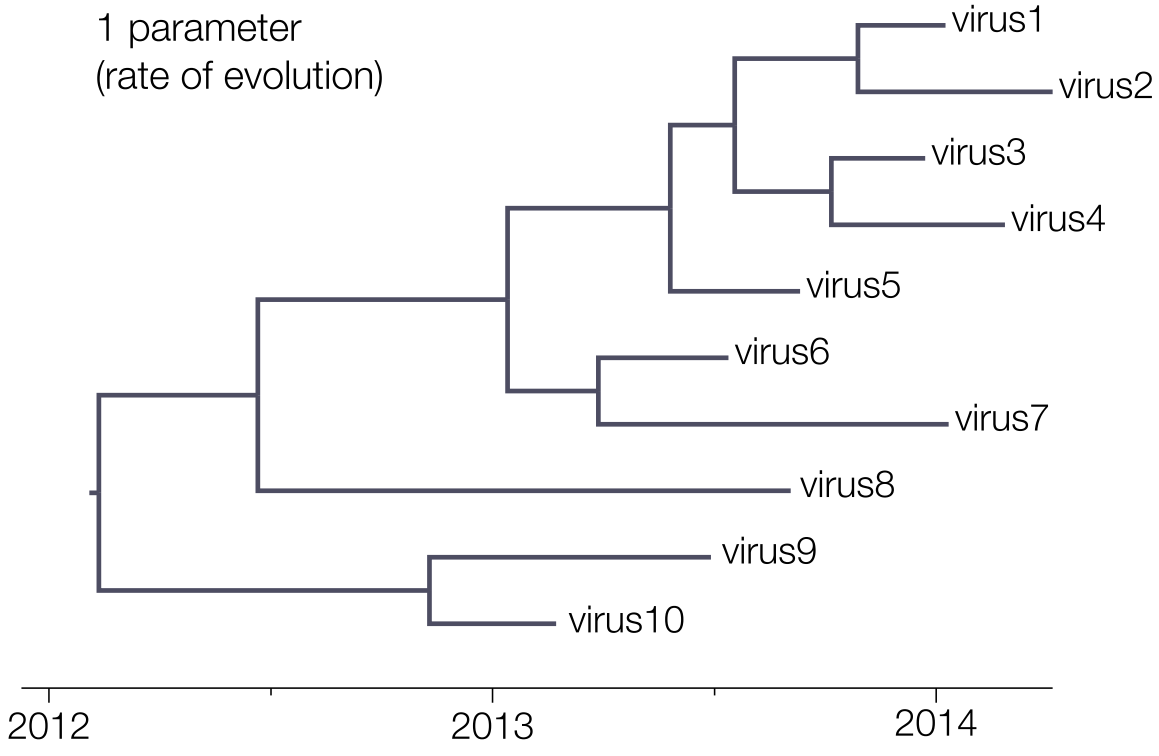
- the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum.
- dominate phylogenetic inference
- but both are biologically unrealistic:
  - the real evolutionary process lies between these two extremes
  - model misspecification can produce positively misleading results

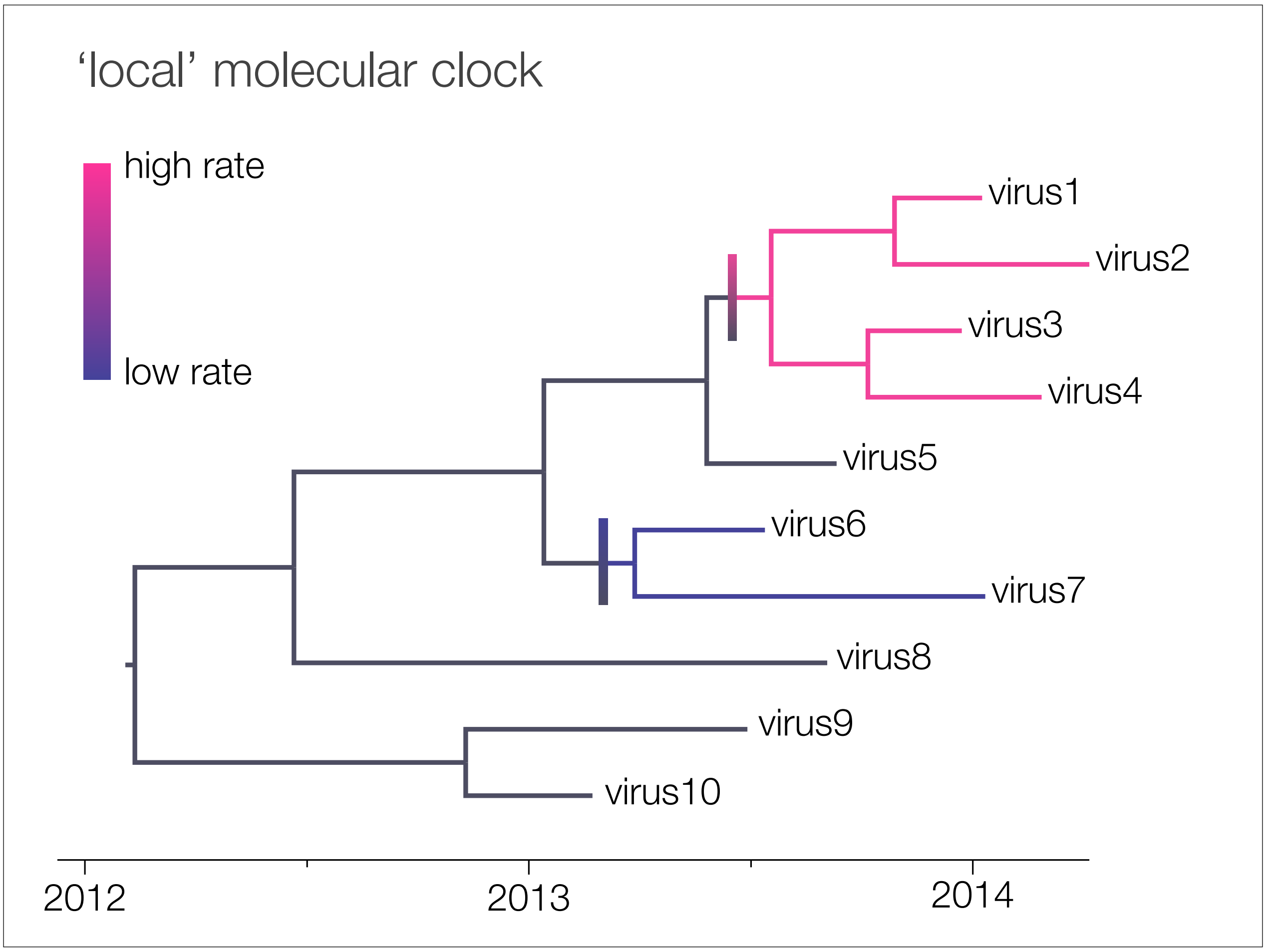


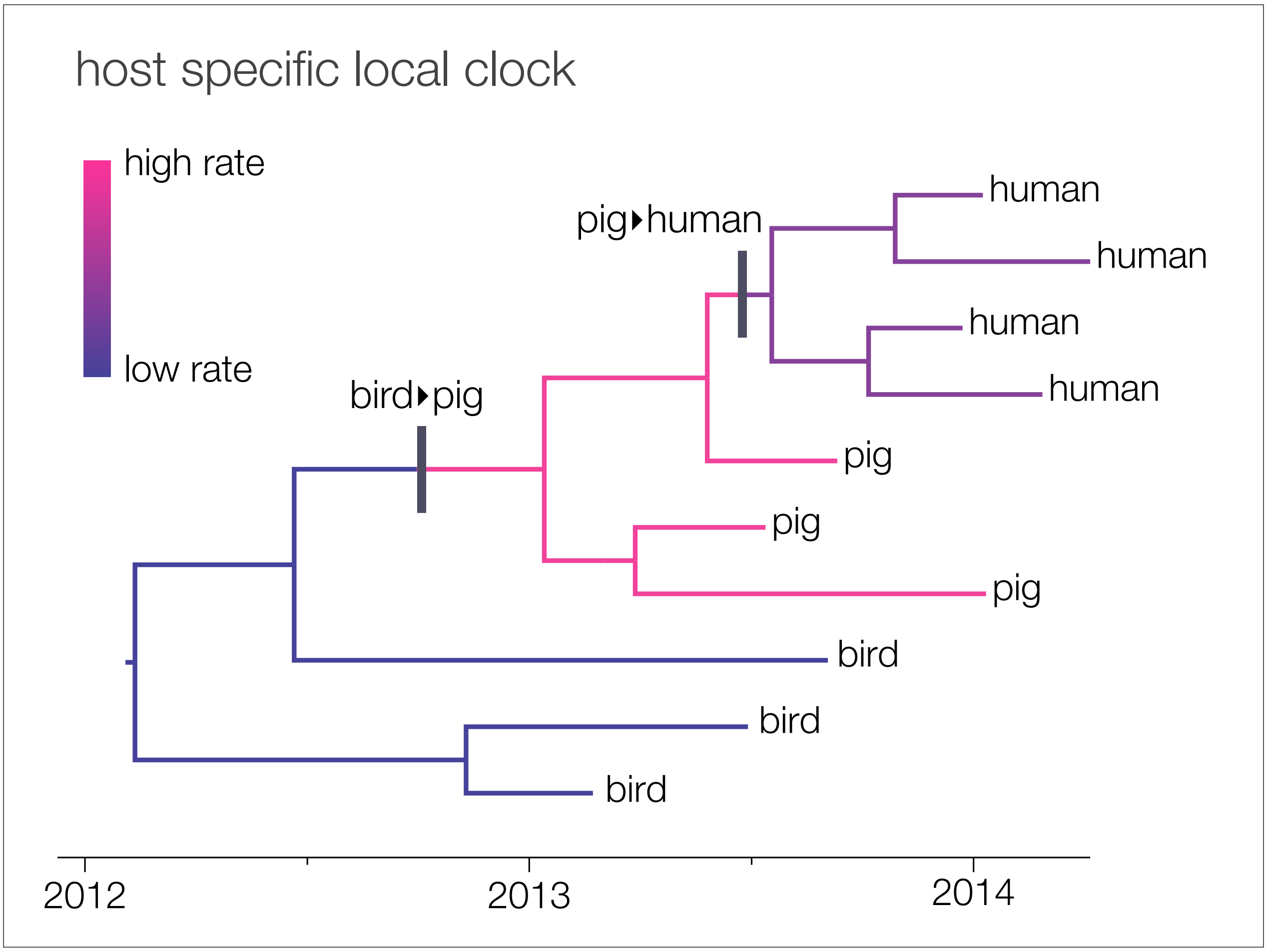
- Pybus (2006) *Genome Biol*. **4**, e151

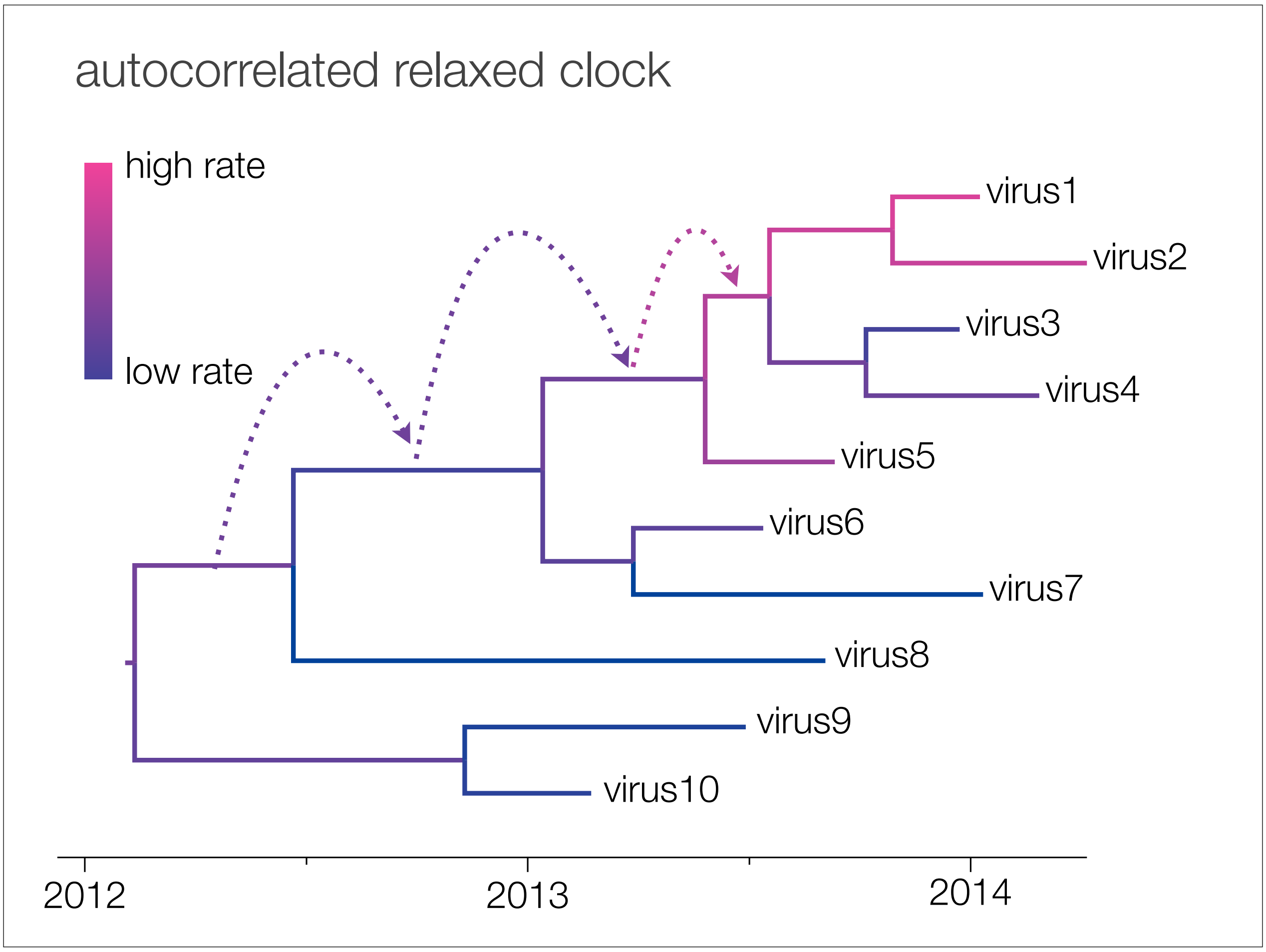
‘strict’ molecular clock

1 parameter
(rate of evolution)

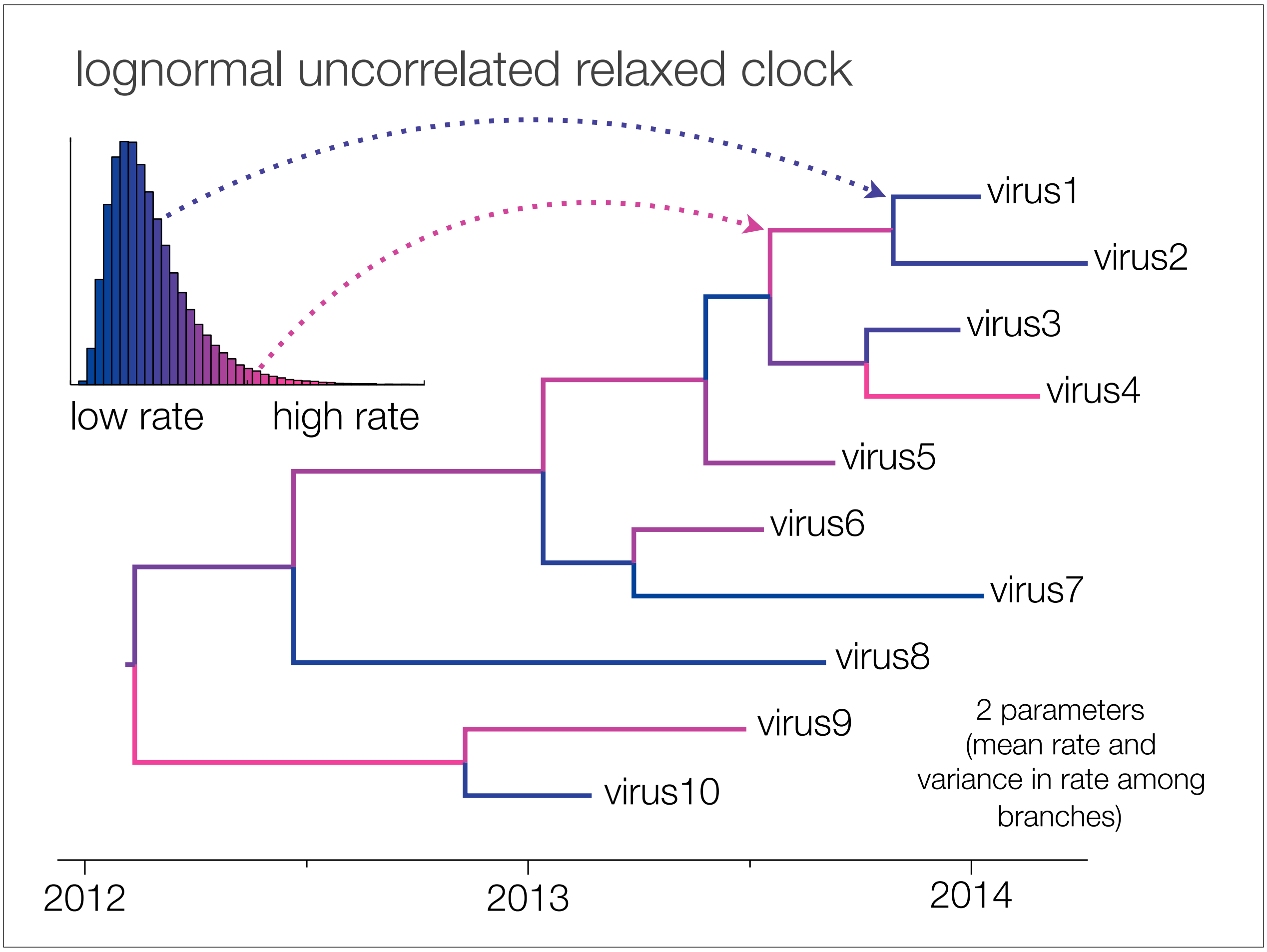



‘local’ molecular clock
high rate
low rate
virus1
virus2
virus3
virus4
virus5
virus6
virus7
virus8
virus9
virus10
2012
2013
2014


host specific local clock
high rate
low rate
pig▸human
bird▸pig
human
human
human
human
pig
pig
pig
bird
bird
bird
2012
2013
2014



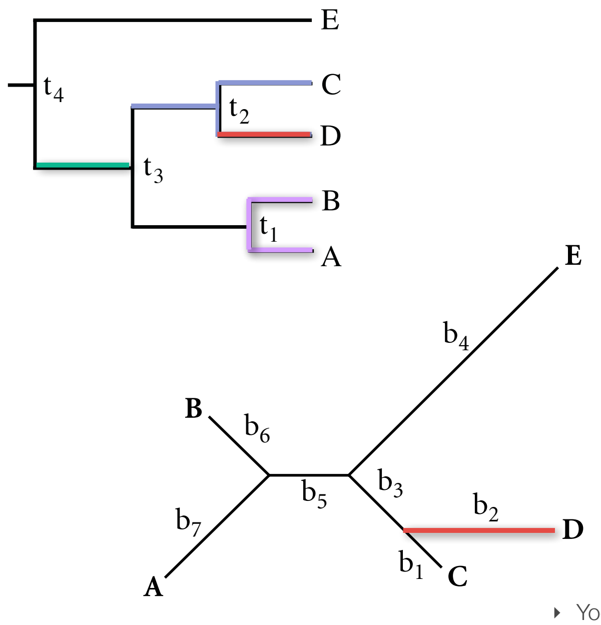
autocorrelated relaxed clock

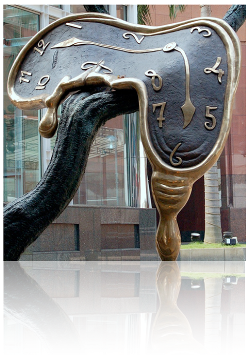


lognormal uncorrelated relaxed clock

2 parameters (mean rate and variance in rate among branches)

Relaxed clocks: (1) local molecular clocks

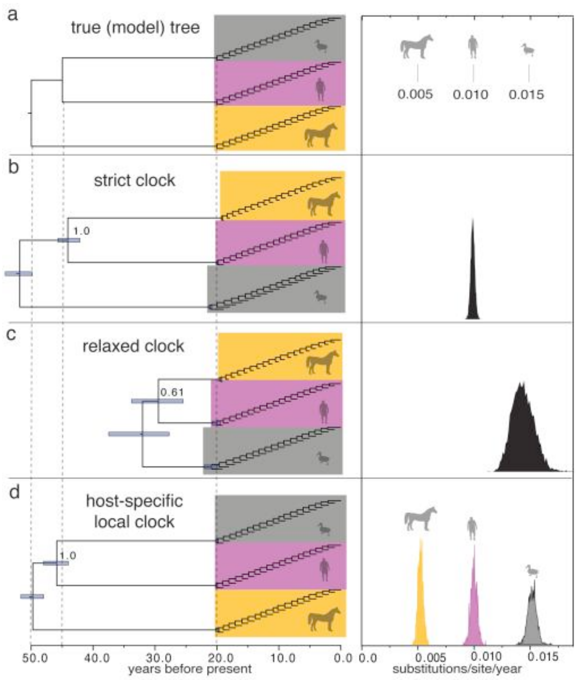


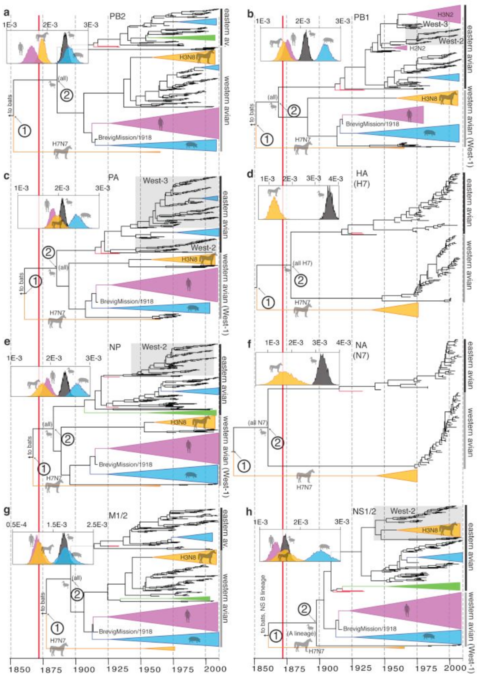


- specify H0 beforehand
- problem of identifiability

- Yoder and Yang (2000) Mol Biol & Evol **17**: 1081-1090.

Bayesian local clocks

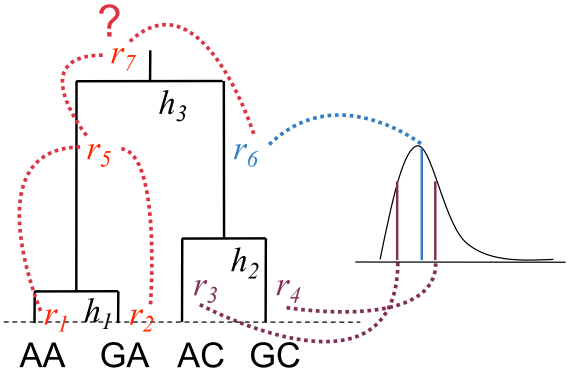




Worobey et al., Nature, 2014; 508(7495): 254–257

Autocorrelated relaxed clocks

- rates for each branch are drawn from a distribution centered on the rate of the ancestor
  - but what is the rate at the root?
  - A prior degree of autocorrelation?
  - not currently possible to do phylogenetic inference

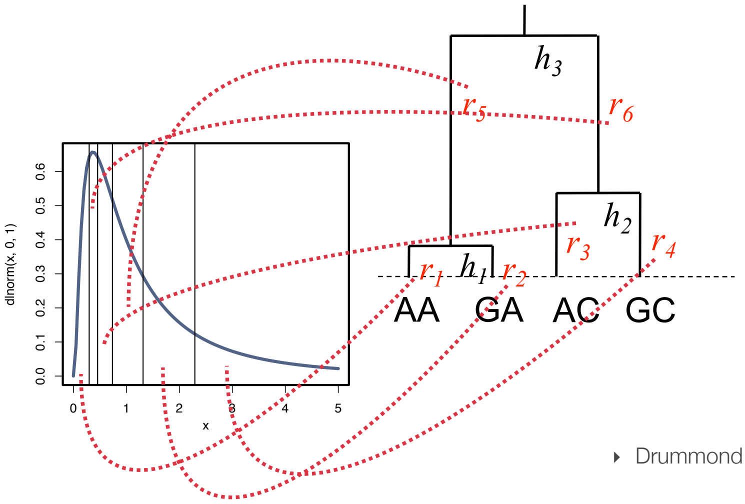


$$r_i \sim LogNormal(r_{A(i)}, \overset{?}{\sigma}^2 \Delta t_i)$$

- e.g., Thorne JL, Kishino H, Painter IS (1998) Mol Biol & Evol **15**: 1647-1657.

Uncorrelated relaxed clocks

- rates for each branch are drawn independently from an identical distribution:



$$r \sim Exp(\lambda)$$

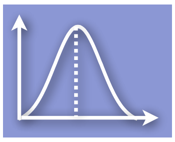
$$r \sim LogNormal(\mu, \sigma^2)$$

$$r \sim Gamma(\alpha, \beta)$$

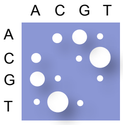
- Drummond et al. (2006) Plos Biology **4**: e88.

Bayesian evolutionary analysis sampling trees

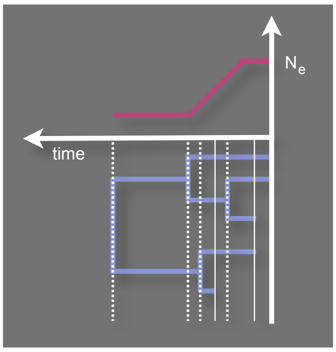
- Given sequence data that is temporally spaced estimate true values of:
  - substitution parameters (μ and Q)
  - ancestral genealogy (g = E_g , t_Y)
    - tree topology
    - dates of divergence
  - population history (θ)



μ



Q



- Bayesian inference

$$P(g,\mu,\theta,Q|D)= \frac{1}{Z} \Pr\{D|\boxed{g,\mu,Q}\}f_g(g|\theta)\underline{f_\mu(\mu)f_\theta(\theta)f_Q(Q)}$$

“relaxed phylogenetics and dating with confidence”

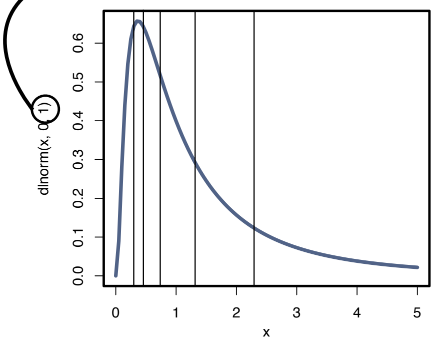
$$t = \{ t_1, t_2 \dots , t_{2n-1} \}$$

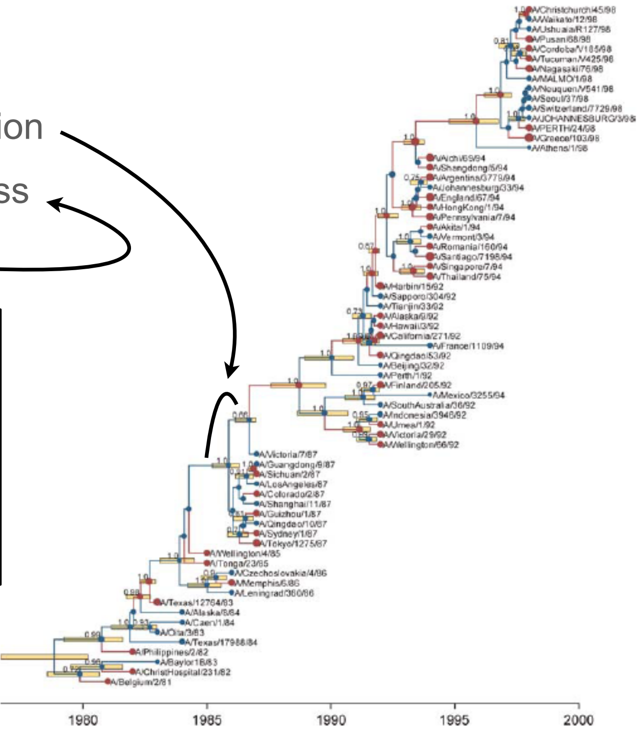
$$R = \{ r_1, r_2 \dots , r_{2n-1} \}$$

$$f(R|g) = f(R) = \prod_{i=1} \lambda e^{-\lambda r_i}$$

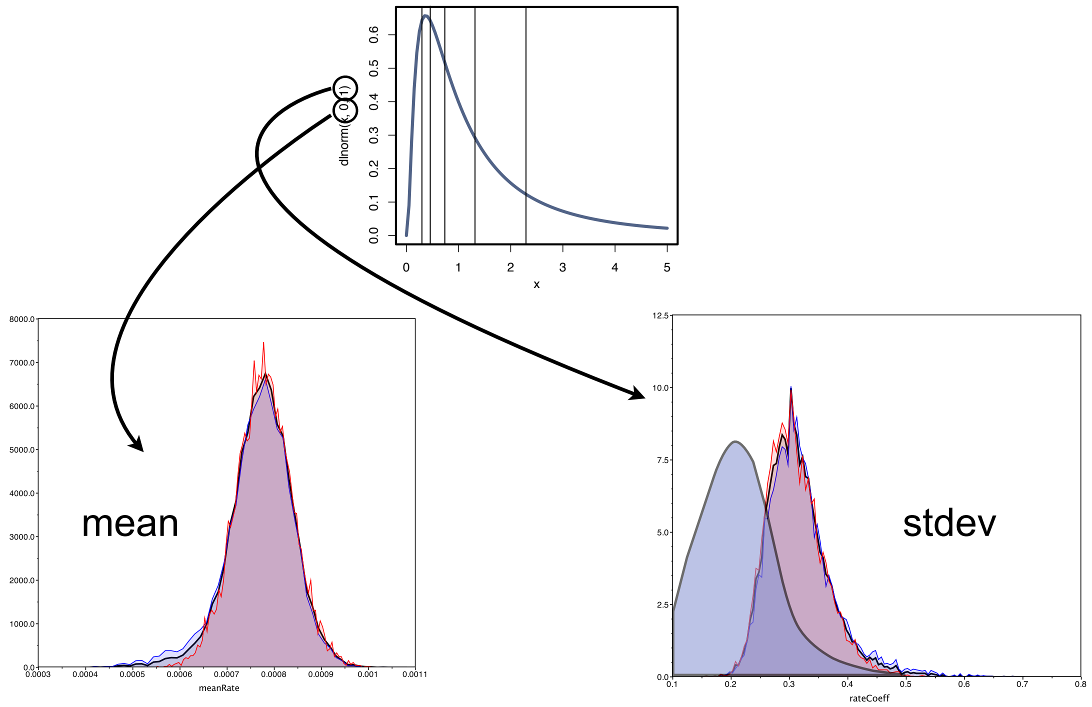
Uncorrelated relaxed clocks: example

- Phylogenetic inference
- measuring autocorrelation
- measuring clock-likeness





Evaluating clock-like behaviour?



Model testing using Bayes factors

- Bayes factor $$B_{01} = \frac{p(Y|M_1)}{p(Y|M_0)}$$

- when two models M_0 and M_1 are being compared, one defines the Bayes factor in favor of M_1 over M_0 as the ratio of their respective marginal likelihoods

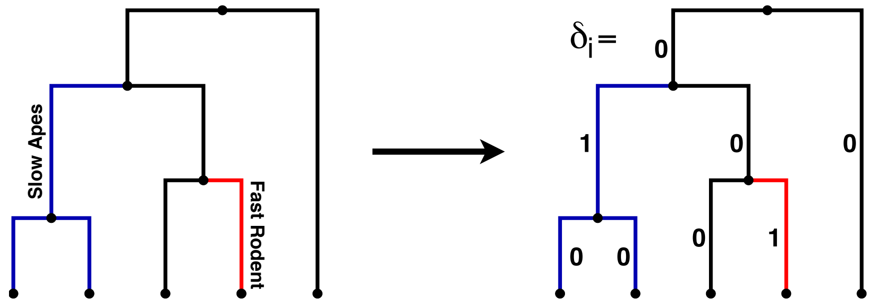
- When there are unknown parameters, the Bayes Factor B_{01} has in a sense the form of a *likelihood ratio*

→ Guy

Random local clocks

- ➡ critics on the local clocks
  - specify H0 a priori
  - problem of identifiability
- ➡ critics on the uncorrelated relaxed clocks
  - Rate changes do not necessarily occur regularly or on every branch
  - Small number of significant changes

So, can we handle the uncertainty in the number and locations of a small number of local clocks?

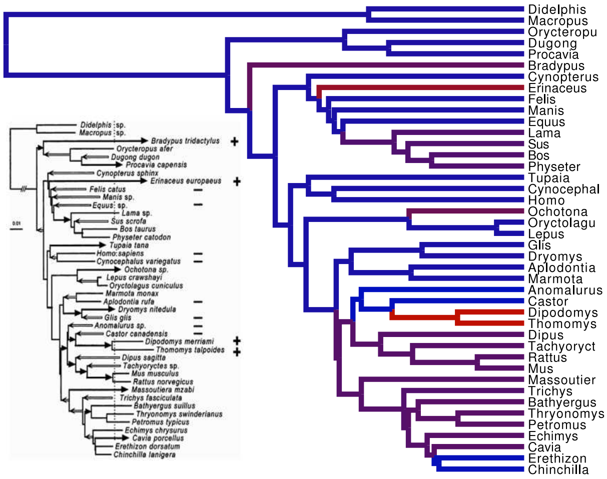


- three local clocks
- two rate changes

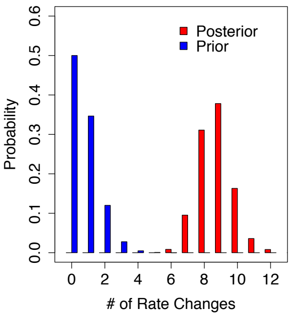
➡ How to explore 2^{2n-2} clock models?

Random local clocks

➡ Using Bayesian stochastic search variable selection: formulate a prior that such that many rate changes (indicators) are 0 but allow the data to determine which ones are required to explain (most of the) rate variation using MCMC



- ➡ Three mtDNA nuclear genes from 42 mammals (Douzery, 2003)
- ➡ 5-12 local clocks



Drummond and Suchard, 2010.

Random local clocks

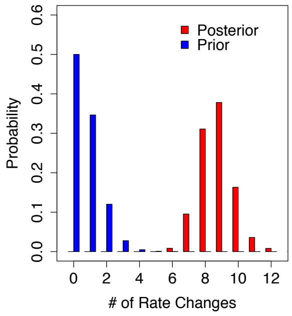
➡ Testing whether a branch accommodates a rate change using Bayes factors

- Data D is assumed to have been arisen under one of two models, or one of two hypotheses H_1 and H_2.

$$\mathrm{pr}(H_k|\mathbf{D}) = \frac{\mathrm{pr}(\mathbf{D}|H_k)\mathrm{pr}(H_k)}{\mathrm{pr}(\mathbf{D}|H_1)\mathrm{pr}(H_1) + \mathrm{pr}(\mathbf{D}|H_2)\mathrm{pr}(H_2)}$$

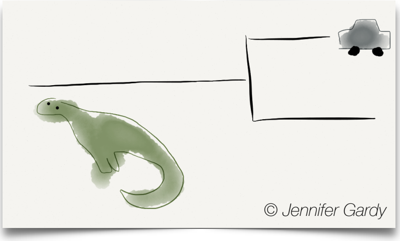
so that

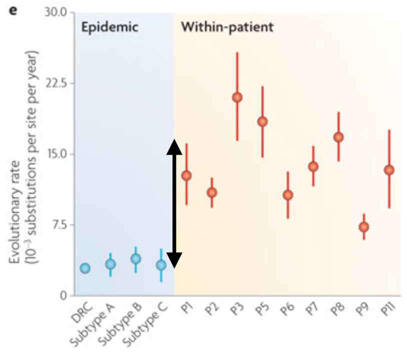
$$\underbrace{\boxed{\frac{\mathrm{pr}(H_1|\mathbf{D})}{\mathrm{pr}(H_2|\mathbf{D})}}}_{\text{posterior odds}} = \underbrace{\boxed{\frac{\mathrm{pr}(\mathbf{D}|H_1)}{\mathrm{pr}(\mathbf{D}|H_2)}}}_{\text{Bayes factor}}\underbrace{\boxed{\frac{\mathrm{pr}(H_1)}{\mathrm{pr}(H_2)}}}_{\text{prior odds}}$$



- Prior probabilities pr(H_1) and pr(H_2) = 1 - pr(H_1). Posterior probabilities pr(H_1 |D) and pr(H_2 |D) = 1 - pr(H_1 |D)

Extensions for testing evolutionary rate hypotheses


© Jennifer Gardy



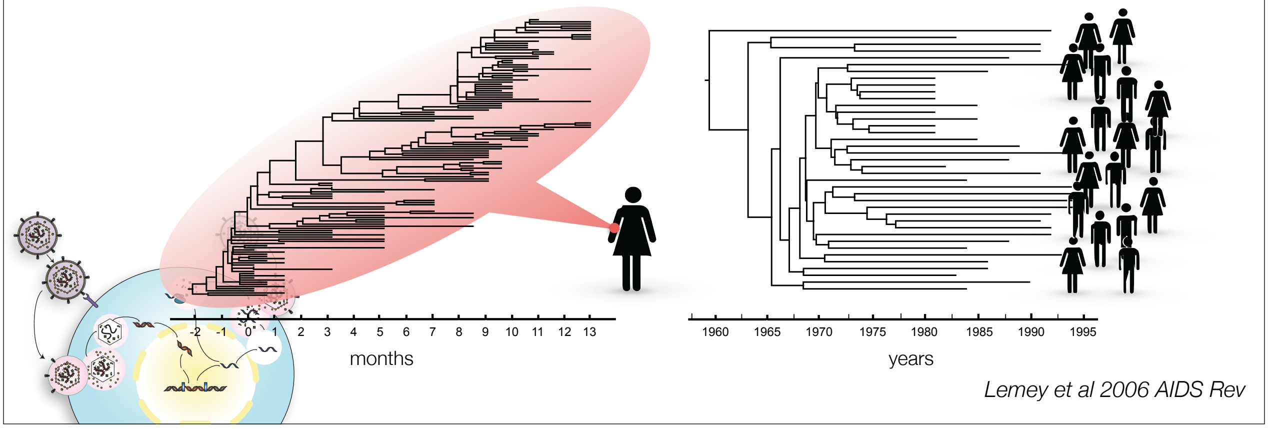
Pybus and Rambaut, NGR, 2009

PROCEEDINGS
OF
THE ROYAL
SOCIETY B

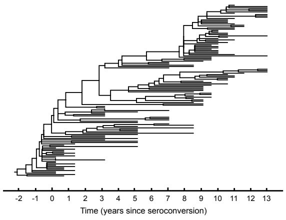
Proc. R. Soc. B (2012) **279**, 3367–3375
doi:10.1098/rspb.2012.0595
Published online 16 May 2012

New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels

Katrina A. Lythgoe* and Christophe Fraser



Lemey et al 2006 AIDS Rev

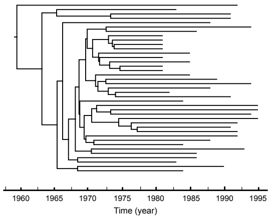


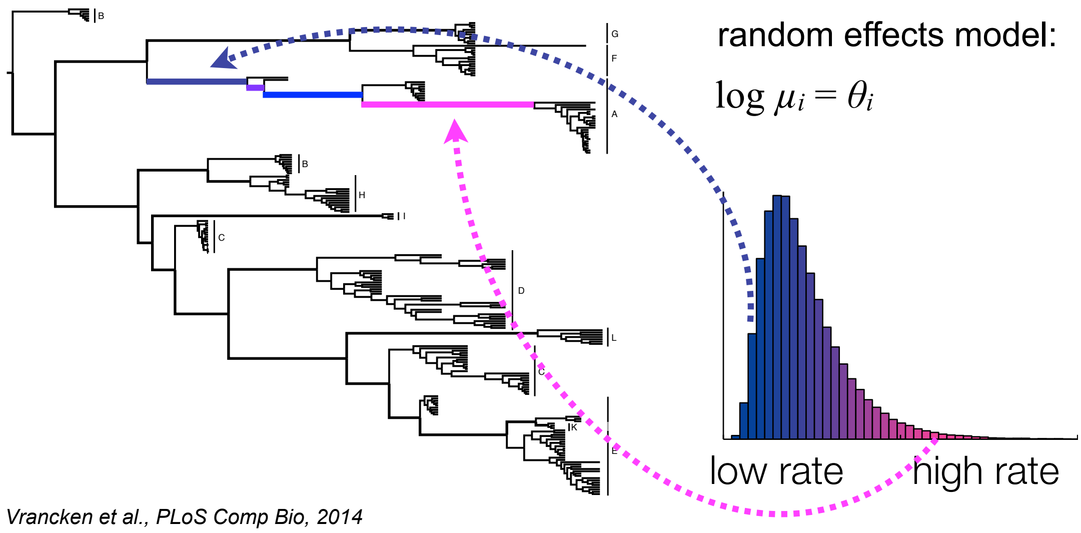
PROCEEDINGS
OF
THE ROYAL
SOCIETY B

Proc. R. Soc. B (2012) **279**, 3367–3375
doi:10.1098/rspb.2012.0595
Published online 16 May 2012

New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels

Katrina A. Lythgoe* and Christophe Fraser

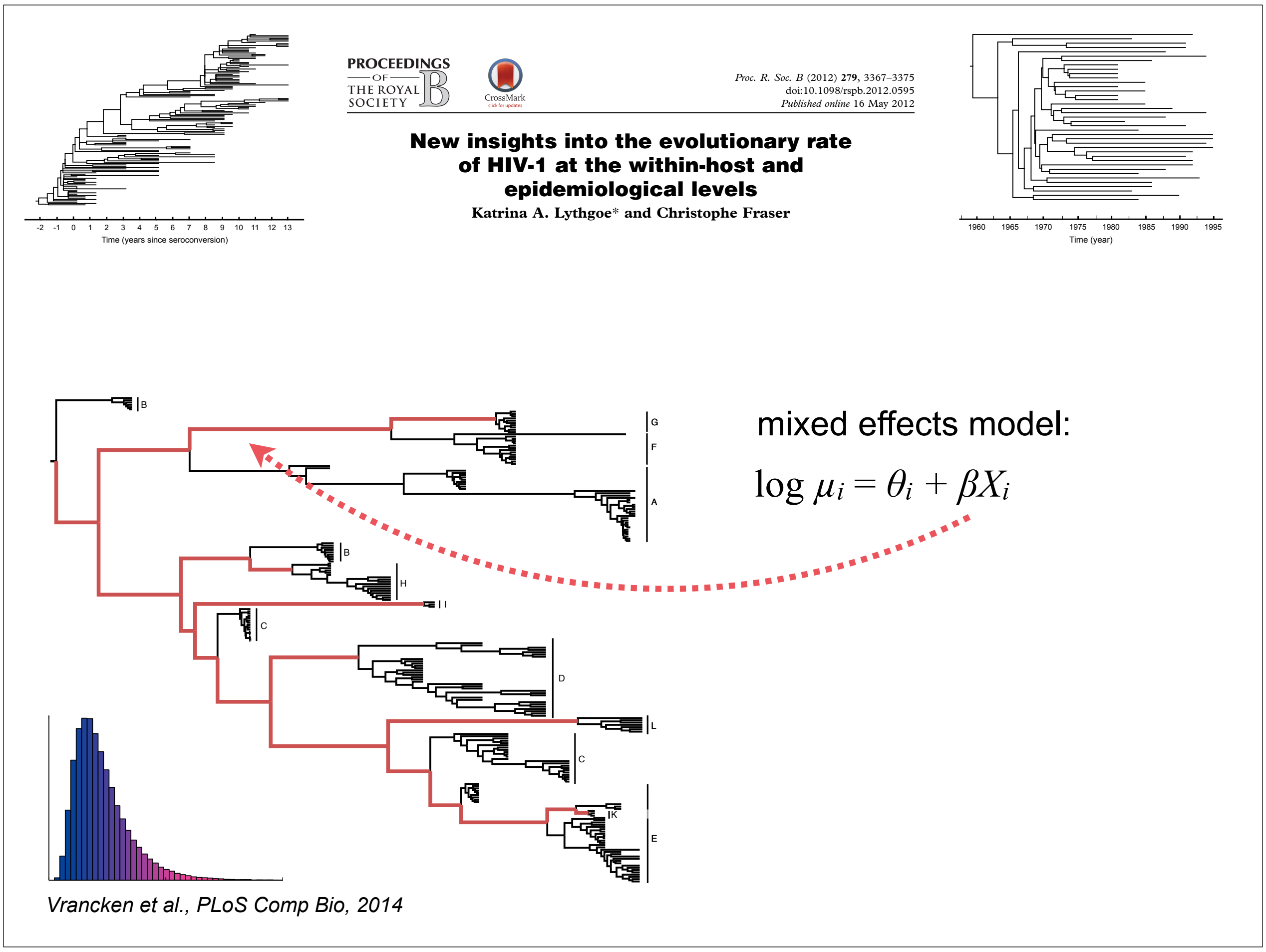




random effects model:

$$\log \mu_i = \theta_i$$

Vrancken et al., PLoS Comp Bio, 2014



PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) 279, 3367–3375
doi:10.1098/rspb.2012.0595
Published online 16 May 2012

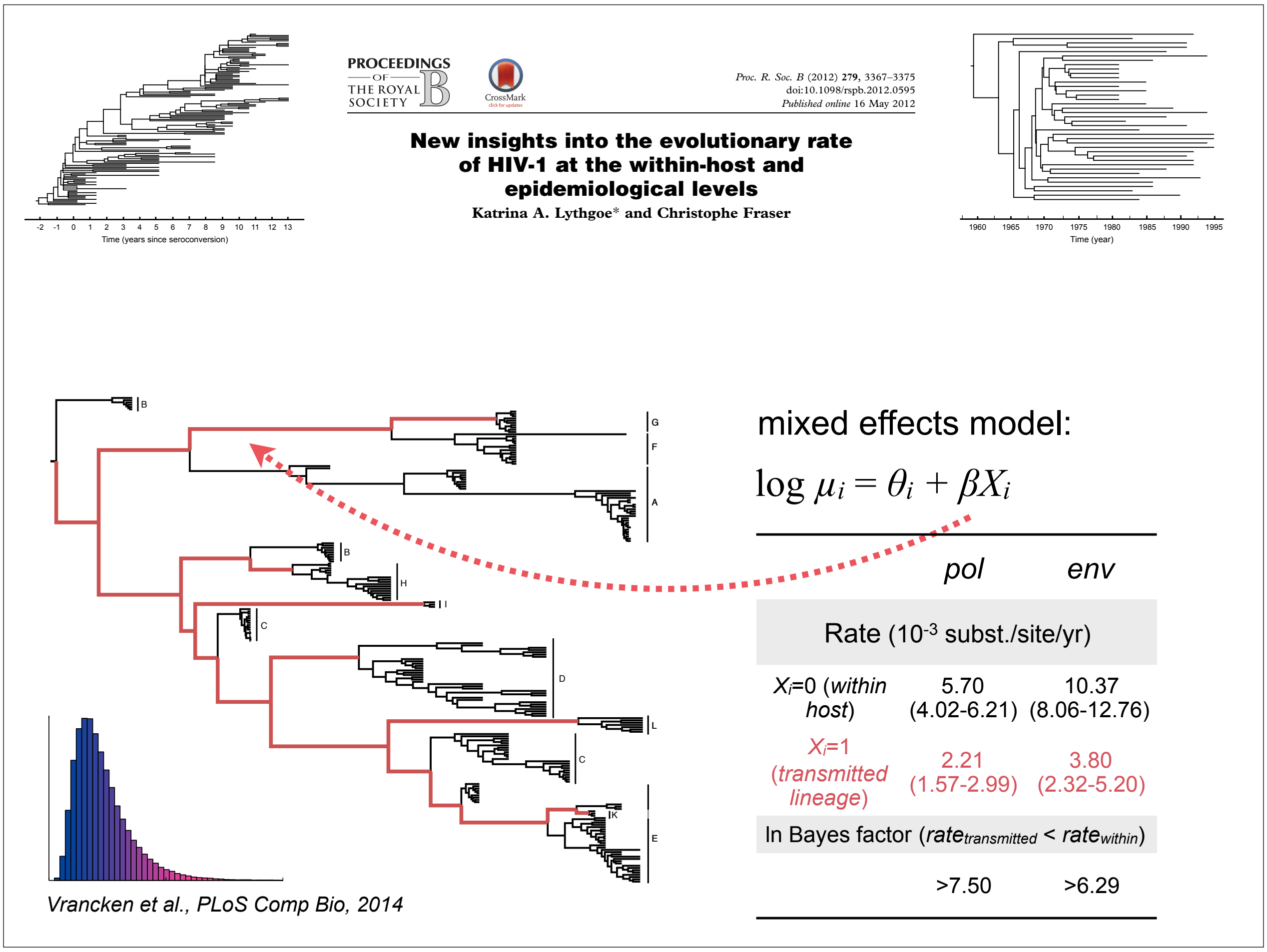
New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels

Katrina A. Lythgoe* and Christophe Fraser

mixed effects model:

$$\log \mu_i = \theta_i + \beta X_i$$

Vrancken et al., PLoS Comp Bio, 2014



PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) 279, 3367–3375
doi:10.1098/rspb.2012.0595
Published online 16 May 2012

New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels

Katrina A. Lythgoe* and Christophe Fraser

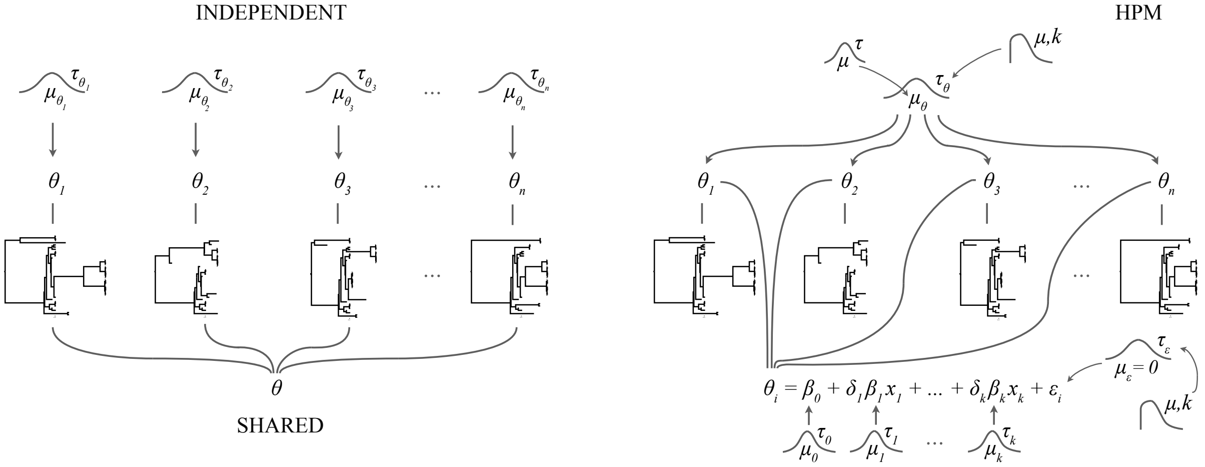
mixed effects model:

$$\log \mu_i = \theta_i + \beta X_i$$

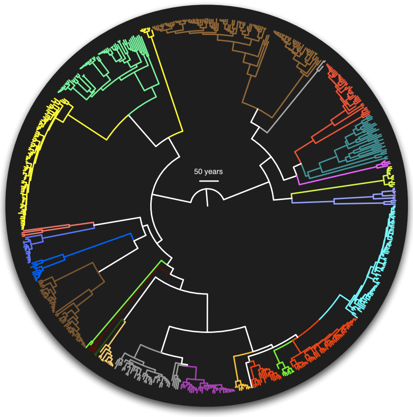
	pol	*env*
Rate (10^{-3} subst./site/yr)		
X_i=0 (*within host*)	5.70 (4.02-6.21)	10.37 (8.06-12.76)
X_i=1 (*transmitted lineage*)	2.21 (1.57-2.99)	3.80 (2.32-5.20)
ln Bayes factor (*rate*$_{transmitted}$ < *rate*$_{within}$)		
	>7.50	>6.29

Vrancken et al., PLoS Comp Bio, 2014

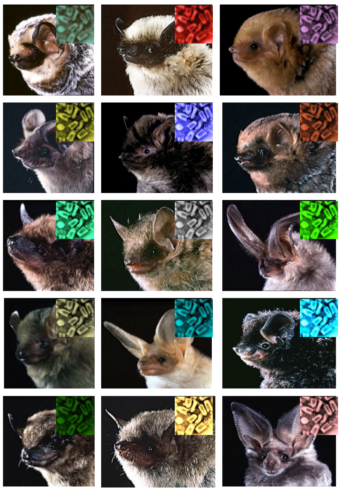
Hierarchical phylogenetic modelling


INDEPENDENT
μ_{θ_1} τ_{θ_1}
μ_{θ_2} τ_{θ_2}
μ_{θ_3} τ_{θ_3}
…
μ_{θ_n} τ_{θ_n}
θ_1 θ_2 θ_3 … θ_n
…
θ
SHARED
HPM
μ τ
μ,k
μ_θ τ_θ
θ_1 θ_2 θ_3 … θ_n
…
τ_ε
$\mu_\varepsilon = 0$
μ,k
$\theta_i = \beta_0 + \delta_1\beta_1x_1 + \ldots + \delta_k\beta_kx_k + \varepsilon_i$
μ_0 τ_0
μ_1 τ_1
…
μ_k τ_k

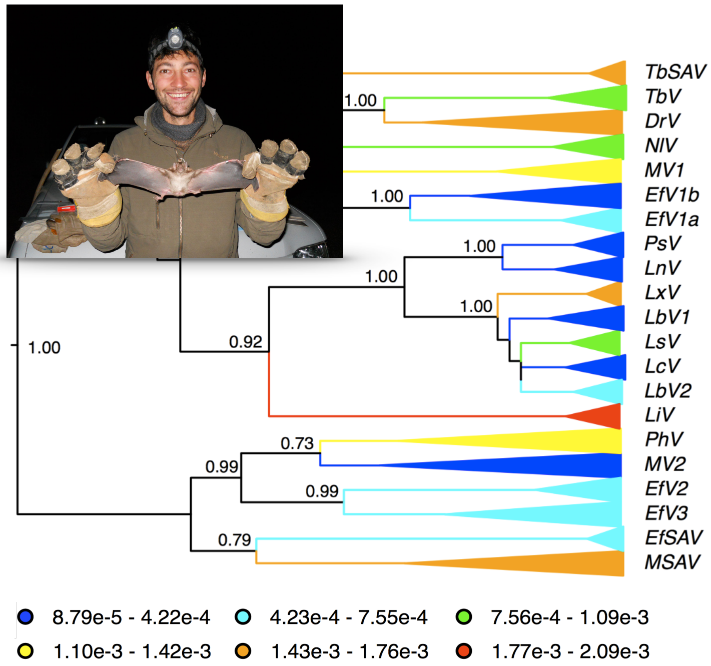
…with fixed effects

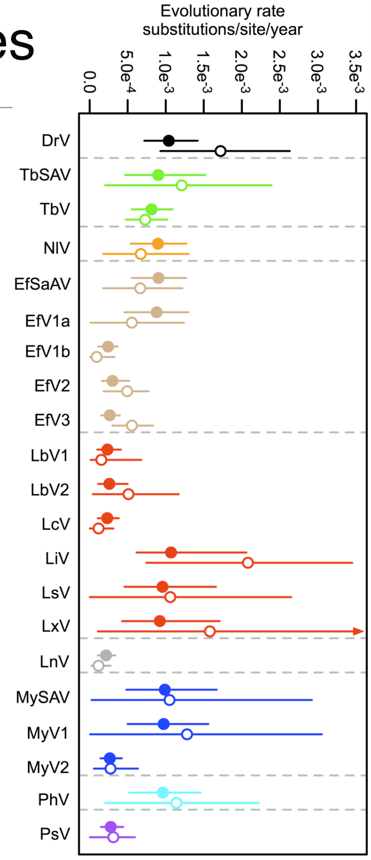


Courtesy of D. Streicker



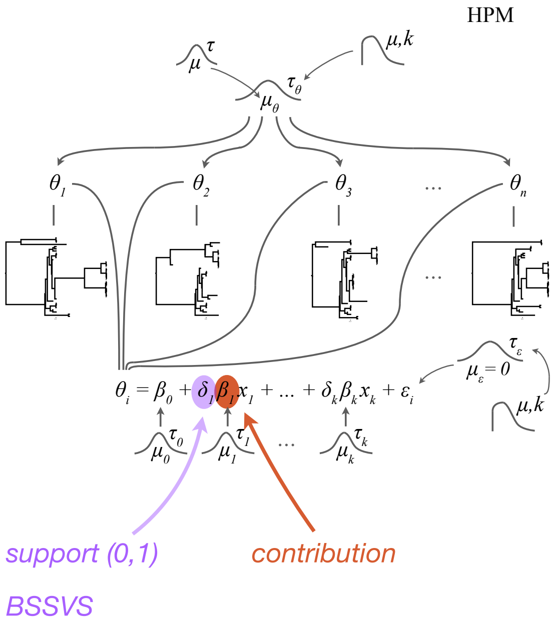
Bat rabies virus evolutionary rates

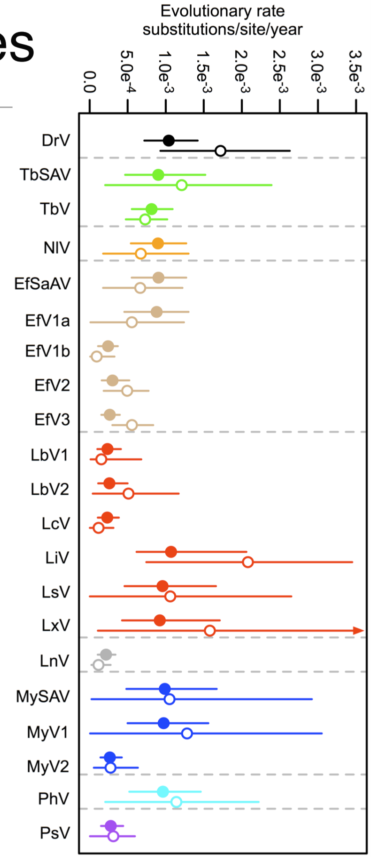




Streicker et al., 2012. *PLoS Pathogens*

Bat rabies virus evolutionary rates

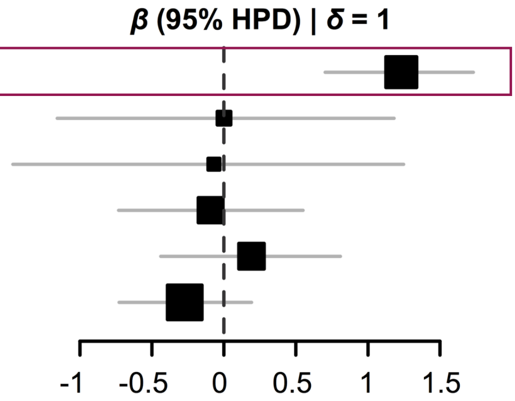


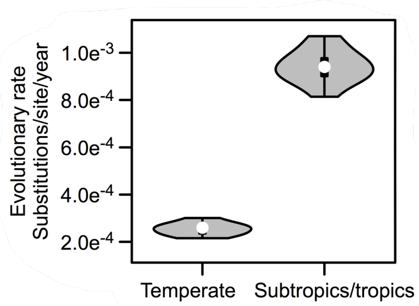


Streicker et al., 2012. *PLoS Pathogens*

Bat rabies virus evolutionary rates

Predictor	Bayes factor
Climate	466.54
Basal metabolic rate	0.82
Torpid metabolic rate	1.00
Coloniality	0.46
Seasonal activity	0.46
Long-distance migration	0.69





Streicker et al., 2012. *PLoS Pathogens*